In the Specification

Please amend the specification at page 5 lines 10-25 as follows:

Homologues in other organisms are available that can be used for comparative sequence analysis. Multiple alignments are performed to study similarities and differences in a group of related sequences. CLUSTAL W is a multiple sequence alignment package available that performs progressive multiple sequence alignments based on the method of Feng and Doolittle, *J. Mol. Evol.* 25:351-360 (1987), the entirety of which is herein incorporated by reference. Each pair of sequences is aligned and the distance between each pair is calculated; from this distance matrix, a guide tree is calculated, and all of the sequences are progressively aligned based on this tree. A feature of the program is its sensitivity to the effect of gaps on the alignment; gap penalties are varied to encourage the insertion of gaps in probable loop regions instead of in the middle of structured regions. Users can specify gap penalties, choose between a number of scoring matrices, or supply their own scoring matrix for both the pairwise alignments and the multiple alignments. CLUSTAL W for UNIX and VMS systems is available at: ~~ftp.ebi.ac.uk~~. ftp-ebi.ac.uk. Another program is MACAW (Schuler *et al., Proteins, Struct. Func. Genet*, 9:180-190 (1991), the entirety of which is herein incorporated by reference, for which both Macintosh and Microsoft Windows versions are available. MACAW uses a graphical interface, provides a choice of several alignment algorithms, and is available by anonymous ftp at: ~~ncbi.nlm.nih.gov~~ ncbi-nlm.nih.gov (directory/pub/macaw).

Please amend the specification at page 8 lines 2-12 as follows:

A characteristic feature of a large scale shotgun sequencing project is that the sequence

data can be processed and assembled into contiguous sequences (contigs), which represent a reconstruction of the original genome sequence for the cloned fragments. Programs are available in the public domain that can analyze the sequence output and assemble the sequences into larger sequence regions representing contiguous sequences of the target genome. Examples of such programs can be found ~~at~~, for example, ~~http://genome.wustl.edu/gsc, http://www.sanger.ac.uk, and http://www.mbt.washington.edu~~ on the World Wide Web at genome.wustl.edu/gsc, www-sanger.ac.uk, and www-mbt.washington.edu. An example of sequence reading program is Phred ~~(http://www.mbt.washington.edu)~~ and can be found of the World Wide Web at (www-mbt.washington.edu). Phred reads DNA sequencer trace data, calls bases, assigns quality values to the bases, and writes the base calls and quality values to output files.

Please amend the specification at page 8 line 13- page 9 line 5 as follows:

The process of assembling DNA sequence fragments generally involves three phases; the overlap phrase, the layout phase and the multi-alignment, or consensus, phase. In the overlap phase, each fragment is compared against every other fragment to determine if they share a common subsequence, an indication that they were potentially sampled from overlapping stretches of the original DNA strand. Pairs of fragments are compared in two ways; 1) with both fragments in the same relative orientation, and 2) with one of the fragments having been reverse complemented. In the layout phase, a series of alternate assemblies or layouts of the fragments based on the pairwise overlaps is generated. A layout specifies the relative locations and orientations of the fragments with respect to each other and is typically visualized as an arrangement of overlapping directed lines, one for each fragment. The general criterion for the

layout phase is to produce plausible assemblies of maximum likelihood. In this manner, it can be determined if there is more than one way to put the pieces together and if different solutions appear equally plausible. In such a case, one would return to the lab and obtain additional information to resolve the ambiguity. The multi-alignment, or consensus, phase uses more information than just the pairwise alignments in the layout. The sequences of all the fragments in a layout are simultaneously aligned, giving a final set of contigs representing regions of the target genome. An example of an assembly program is PHRAP, which can be found on the World Wide Web at ~~http://chimera.biotech.washington.edu/UWGC/tools/phrap.htm~~ chimera.biotech.washington.edu/UWGC/tools/phrap.htm.

Please amend the specification at page 90 lines 12-21 as follows:

PHRED is used to call the bases from the sequence trace files ~~(http://www.mbt.washington.edu)~~ (www-mbt.washington.edu). Phred uses Fourier methods to examine the four base traces in the region surrounding each point in the data set in order to predict a series of evenly spaced predicted locations. That is, it determines where the peaks would be centered if there were no compressions, dropouts, or other factors shifting the peaks from their "true" locations. Next, PHRED examines each trace to find the centers of the actual, or observed peaks and the areas of these peaks relative to their neighbors. The peaks are detected independently along each of the four traces so many peaks overlap. A dynamic programming algorithm is used to match the observed peaks detected in the second step with the predicted peak locations founds in the first step.

Please amend the specification at page 91, lines 1-19 as follows:

Contigs are assembled using PHRAP ~~(http://www.mbt.washington.edu)~~ (www-mbt.washington.edu). Contigs and singletons are interrogated using AAT-NAP and BLASTP. AAT_NAP is a program used for constructing a global alignment of a DNA sequence and a protein sequence (Huang, X. *et al., Genomics* 46:37-45 (1997), the entirety of which is herein incorporated by reference). The alignment model of NAP accommodates introns and frameshifts within codons. The scheme for scoring an alignment has several features that allow NAP to identify the exact locations of introns. A nucleotide insertion gap of length < k is given a linear penalty, and a nucleotide insertion gap of length > k is penalized as a gap of length k, where the value for k is the default value. The NAP program reports the starting and ending coordinates of predicted genes. The input to the NAP program includes the query sequence, the protein database and a coordinate file produced by AAT_EXT (an adapter between a database search program and a sequence alignment program) from the output of AAT_DPS (a program computing high-scoring chains of segment pairs between a query DNA sequence and the public non-redundant protein database from NCBI). The NAP program scans the protein database and finds the protein sequence for each coordinate record. Then for each coordinate record, NAP locates the query region, extends the region in both directions by a certain number of bases, and computes an alignment of the extended region and the protein sequence.